AFRL-RY-WP-TR-2014-0223

DISCOVERY OF DEEP STRUCTURE FROM UNLABELED DATA

Andrew NG and Christopher Manning

The Leland Stanford Junior University

NOVEMBER 2014
Final Report

AIR FORCE RESEARCH LABORATORY
SENSORS DIRECTORATE
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7320
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the USAF 88th Air Base Wing (88 ABW) Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RY-WP-TR-2014-0223 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

*//Signature//

PHILIP D. MUMFORD, PhD
Program Manager
Avionics Vulnerability Mitigation Branch
Spectrum Warfare Division

//Signature//

DAVID HAGSTROM, Chief
Avionics Vulnerability Mitigation Branch
Spectrum Warfare Division

*//Signature//

TODD A. KASTLE, Chief
Spectrum Warfare Division
Sensors Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

*Disseminated copies will show "//signature//" stamped or typed above the signature blocks.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
November 2014	Final	15 March 2010 – 16 June 2014

4. TITLE AND SUBTITLE
DISCOVERY OF DEEP STRUCTURE FROM UNLABELED DATA

5a. CONTRACT NUMBER
FA8650-10-C-7020

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
61101E

6. AUTHOR(S)
Andrew NG and Christopher Manning

5d. PROJECT NUMBER
1000

5e. TASK NUMBER
00

5f. WORK UNIT NUMBER
Y0J2

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The Leland Stanford Junior University
450 Serra Mall
Stanford, CA 94305

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory
Sensors Directorate
Wright-Patterson Air Force Base, OH 45433-7320
Air Force Materiel Command
United States Air Force

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
AFRL/RYW

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RY-WP-TR-2014-0223

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
PAO Case Number 88ABW-2014-4872, Clearance Date 23 October 2014. Report contains color.

14. ABSTRACT
This research project addressed the problem of learning useful "deep" representations from unlabeled data. The major goal was to innovate new unsupervised deep learning algorithms capable of learning important semantic structure in the input data in a domain general way. At the conclusion of this project, these goals stand fulfilled. The lab produced a variety of new and influential learning algorithms including Independent Subspace Analysis (ISA); Reconstruction Independent Components Analysis (RICA); recursive neural networks; and recursive tensor networks, among others. These algorithms have posted state-of-the-art results across a number of domains and tasks, and have had impact on both academia and industry.

15. SUBJECT TERMS
unsupervised learning, reduced boltzman machine, recursive neural networks

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT:	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON (Monitor)
a. REPORT	b. ABSTRACT	c. THIS PAGE			Philip Mumford
Unclassified	Unclassified	Unclassified	SAR	24	19b. TELEPHONE NUMBER *(Include Area Code)* N/A

TABLE OF CONTENTS

1.0 SUMMARY

This research project addressed the problem of learning useful "deep" representations from unlabeled data. The major goal was to innovate new unsupervised deep learning algorithms capable of learning important semantic structure in the input data in a domain general way. At the conclusion of this project, these goals stand fulfilled. The lab produced a variety of new and influential learning algorithms including Independent Subspace Analysis (ISA); Reconstruction Independent Components Analysis (RICA); recursive neural networks; and recursive tensor networks, among others. These algorithms have posted state-of-the-art results across a number of domains and tasks, and have had impact on both academia and industry.

2.0 INTRODUCTION

This research project addressed the problem of learning useful "deep" representations from unlabeled data. How can we learn succinct, higher-level representations of a variety of input data? The major goal was to innovate new unsupervised deep learning algorithms capable, ultimately, of learning important semantic structure in the input data in a domain general way. As evidence of this ability, we applied the models to video, text, and other modalities; and even to multiple modalities simultaneously. At the outset of the project, the promise of these algorithms was to allow general-purpose machine learning to be much more easily applied to problems in vision, audio understanding, text understanding, sensor understanding, and other problems, and achieve superior performance while requiring significantly less hand tuning. Instead of needing hand-engineered features developed through years of research, a generic deep learner would be able to achieve superior performance using large amounts of unlabeled and labeled data.

Our approach relied on “deep learning” techniques, which learn many stages of nonlinear transformations. The learning process can be fully supervised, but can also leverage unsupervised data in a “pretraining” stage. Our algorithms were deployed on commodity hardware. A key method in our approach was to scale these algorithms to much greater sizes and apply them to much larger datasets than had previously been attempted. This required better algorithms specifically adapted to this scale, as well as large clusters of commodity computers and GPU processors. To evaluate the unsupervised learning component of the algorithms (which has become of less importance in the era of “big data”, see Discussion and Conclusions) we deployed advanced visualization techniques to understand invariance to transformations; compared learned representations to those in biological visual, auditory, and somatosensory cortex; and ran numerous control experiments investigating the impact of architecture, learning algorithms, and encoding methods. These methods are further discussed in Section 3.0.

In Section 4.1 we discuss results from the variety of new and influential learning algorithms we developed, including ISA; RICA; recursive neural networks; and recursive tensor networks, among others. These have bested state-of-the-art results across a large range of domains including image recognition, auditory phoneme recognition, image segmentation, parsing, sentiment classification, knowledge base reasoning, video activity recognition, and multimodal audio-video and text-image learning. The breadth of tasks on which these techniques have been successful is unusual, and points to the largely domain general nature of these machine learning methods. As a mark of the impact of this work, the academic papers reporting our results have been cited many hundreds of times, and the methods have been reported in a variety of popular press venues including Wired and the New York Times.

A key and distinctive feature of this grant was its focus on scaling: from the outset, our lab intended to push towards very large models using a combination of COTS hardware and GPU acceleration. Deep learning models do not naturally parallelize, so this scaling has required the introduction of a number of new algorithmic techniques. The hypothesis was that a simple algorithm running at huge scale would do better than a complex

algorithm that could only be run on smaller datasets. This idea has proven deeply impactful, from initial work showing that simple feature learning algorithms, if scaled massively, could substantially beat state-of-the-art algorithms, through multiple 10x size increases in the largest-ever trained networks, each associated with new record results. Indeed even in the short span of this grant, this idea has crossed into industry (based on our lab's work at the Google Brain project), where massive deep networks are now deployed in products in a number of top tech companies. Particularly in industry, where "big data" has generated labeled datasets of previously unthinkable size, scaling has proven to be a central breakthrough emerging from this work. Section 4.2 outlines our results obtained through scaling.

To guide the development of algorithms, we also embarked on a number of theory-focused projects aimed at a greater understanding of deep learning algorithms. Efforts in this direction have ranged from Hessian-based invariance visualizations, to analytic investigations of the impact of architectural choices on selectivity and invariance. Our lab reported the first extensive quantitative comparison of learned representations from a variety of algorithms to those in biological primary visual, auditory, and somatosensory cortices. We also conducted a highly impactful study separating the effect of learning algorithm from encoding architecture in classification performance. A central theme emerging repeatedly from this work is the ultimate similarity of different unsupervised learning algorithms, despite widely differing formal justifications. Whereas it had previously been thought that differences in algorithm performance were largely attributable to differences in the learning algorithms (e.g., learning probabilistic generative models in RBMs; learning independent components in ICA; or learning cluster centroids in K-means), our work has shown that in fact 1) all formalisms match biological data equally well and 2) all formalisms produce learned filters supporting equal classification performance. Differences in classification performance arise, instead, from differences in how these features are encoded—the architecture of the system—rather than differences in the filters arising from the learning process. For example, RBMs use filters in a simple linear map followed by sigmoid nonlinearity, while sparse coding uses filters in an inference step that computes sparse activities—and this inference step—not differences in the learned filters—is primarily responsible for the better classification performance of sparse coding. As one particularly extreme version of these results, we showed that even completely random filters could support near state-of-the-art classification results when embedded in the right architecture. These insights further emphasize our focus on scaling, as the training algorithm has proven to be less important than other factors. Further details on our theoretical results are given in section 4.3.

Section 5.0 sets out the conclusions arising from the research effort, discusses aspects of the original proposal that have changed over the course of the grant, and describes ongoing derivative work.

3.0 METHODS, ASSUMPTIONS, AND PROCEDURES

Deep, layered, compositional representations are the key conceptual method underlying the variety of algorithms and methods we investigated in this grant. Deep learning algorithms build a feature hierarchy of gradually increasing complexity that can support a great variety of applications. Our methods can be divided into efforts aimed at new algorithms, larger scale, and greater understanding (theory).

3.1 Novel algorithms

We have introduced a variety of new models over the course of this grant, listed below. While containing novel aspects, all of these models are extensions of (and arose in dialogue with) prior work by a variety of groups. The primary evaluation for our new methods is classification performance on a practical task (or tasks) of interest. In keeping with the domain general promise of deep learning algorithms, our applications have spanned domains from image recognition to sentiment analysis, speech recognition, and natural language processing, surpassing the previous state-of-the-art in all of these areas.

- **Tiled convolutional neural networks [1]**, which relax the rigidly enforced translation invariance of the popular convolutional neural network model
- **Spatio-temporal Independent Subspace Analysis [2]** which extends ISA to model video input and complex invariances
- **Deep energy models [3]**, which provide a formalism for probabilistic training of deep layered networks
- **Recursive neural networks [4, 5]**, in which deep networks with tree structures are recursively built up for each input example, to capture the recursion present in, eg natural language
- **Recursive autoencoders [6]**, which extend supervised recursive neural networks to the unsupervised setting
- **Dynamic pooling [7],** which extends the commonly used pooling over a fixed neighborhood to a dynamically-sized neighborhood for use with variable length input (eg, natural language)
- **Similarity-based receptive field selection [8]**, an algorithm that learns local receptive fields based on pairwise similarity, replacing the typically hand-crafted local receptive field structure used in most algorithms.
- **Reconstruction Cost Independent Component Analysis [9]**, in which the hard orthonormality of standard ICA is replaced by a soft orthogonalization based on minimizing reconstruction error
- **Sparse filtering [10]**, a simplified feature learning algorithm with only a single hyperparameter
- **Matrix-vector recursive neural networks [11],** which extend recursive neural networks to include both the standard vector encoding
- **Convolutional recursive neural networks [12]**, which combine convolution and recursion to achieve translation invariance and model recursive structure in, eg, images
- **Neural tensor networks [13],** which further generalizes matrix-vector neural

networks

3.2 Methods for deep learning at massive scale

A key focus of the grant is on scaling deep learning methods to much larger models and datasets. To achieve this, we developed novel methods of parallelizing algorithms on commercial off the shelf hardware, including large clusters of CPUs, as well as GPU hardware [14]. We developed algorithms and optimization methods specifically aimed at large scale, including using K-means for unsupervised learning and removing the need for whitening at each layer via soft orthogonalization [9, 15, 16, 17].

3.3 Theory and control experimental methods

While the primary evaluation of our novel methods is classification performance on a task of interest, we have also sought deeper understanding of deep learning algorithms. Towards this end we have occasionally employed more theoretical methods, including mathematical proofs of properties of architectures (eg, provable translation invariance and orientation selectivity in convolutional square pooling architectures with random weights [18]). We have also developed a novel test of the multimodal abilities of deep learning algorithms based on comparisons to neuroscientific data from mammalian visual, auditory, and somatosensory cortices [19]. We also directly compared the performance of different optimization techniques in the context of large scale models [17]. Finally, we developed control experiments to tease apart the relative contribution of different *unsupervised learning* algorithms (that generate a set of weight vectors in a deep hierarchy) as compared to *encoding* algorithms (that use a set of weight vectors to encode a new input example) [15].

4.0 RESULTS AND DISCUSSION

Our novel algorithms have obtained good performance in a broad range of tasks, which we detail in Section 4.1. We describe our performance gains obtained through achieving greater scale in Section 4.2. Finally Section 4.3 presents results from our theoretical analyses.

4.1 Performance of novel algorithms

In the ensuing, we divide our results by modality, first discussing visual applications, then natural language processing applications, and finally multimodal applications.

4.1.1 Visual object recognition. A central thread of our research has concerned applications to visual object recognition problems. We now describe our results obtained for this class of tasks. We note that our large-scale efforts (often also aimed at object recognition) are described in Section 4.2. Here we describe our novel models that were applied to smaller, research-oriented datasets.

In early work [1], we developed the tiled convolutional neural network model. Convolutional neural networks (CNNs) have been successfully applied to many tasks such as digit and object recognition. Using convolutional (tied) weights significantly reduces the number of parameters that have to be learned, and also allows translational invariance to be hard-coded into the architecture. We considered the problem of learning invariances, rather than relying on hardcoding. We proposed tiled convolution neural networks (Tiled CNNs), which use a regular "tiled" pattern of tied weights that does not require that adjacent hidden units share identical weights, but instead requires only that hidden units k steps away from each other to have tied weights. By pooling over neighboring units, this architecture is able to learn complex invariances (such as scale and rotational invariance) beyond translational invariance. Further, it also enjoys much of CNNs' advantage of having a relatively small number of learned parameters (such as ease of learning and greater scalability). We provided an efficient learning algorithm for Tiled CNNs based on Topographic ICA, and showed that learning complex invariant features allows us to achieve highly competitive results for both the NORB and CIFAR-10 datasets.

Next, we developed Deep Energy Models [3]. Deep generative models with multiple hidden layers have been shown to be able to learn meaningful and compact representations of data. Deep energy models use deep feedforward neural networks to model the energy landscapes that define probabilistic models. We are able to efficiently train all layers of the model simultaneously, allowing the lower layers of the model to adapt to the training of the higher layers, and thereby producing better generative models. We evaluated the generative performance of our models on natural images and demonstrated that this joint training of multiple layers yields qualitative and quantitative improvements over greedy layerwise training. We further generalized our models beyond the commonly used sigmoidal neural networks and showed how a deep extension of the product of Student-t distributions model achieves good generative performance. Finally,

we introduced a discriminative extension of our model and demonstrated that it outperforms other fully-connected models on object recognition on the NORB dataset.

We applied deep learning methods to text detection and character recognition in scene images with unsupervised feature learning [20, 21]. Reading text from photographs is a challenging problem that has received a significant amount of attention. Two key components of most systems are (i) text detection from images and (ii) character recognition, and many recent methods have been proposed to design better feature representations and models for both. We instead applied large-scale deep learning algorithms for learning the features automatically from unlabeled data–and showed that they enable highly effective classifiers for both detection and recognition to be used in a high accuracy end-to-end system. Our method attained exceptional performance, far surpassing the previous state of the art.

Finally, we developed algorithms for processing video data [2]. Much work on action recognition has focused on adapting hand-designed local features, such as SIFT or HOG, from static images to the video domain. We instead used unsupervised feature learning as a way to learn features directly from video data. More specifically, we developed an extension of the Independent Subspace Analysis algorithm to learn invariant spatio-temporal features from unlabeled video data. We discovered that, despite its simplicity, this method performs surprisingly well when combined with deep learning techniques such as stacking and convolution to learn hierarchical representations. By replacing hand-designed features with our learned features, we achieved classification results superior to all previous published results on the Hollywood2, UCF, KTH and YouTube action recognition datasets. On the challenging Hollywood2 and YouTube action datasets we obtain 53.3% and 75.8% respectively, which are approximately 5% better than the current best published results. Importantly, prior methods addressed these different datasets individually, while our approach could support good performance on all of them simultaneously, pointing to the domain-general promise of unsupervised learning methods.

4.1.2 Natural language processing. Another strong thread of our research has been novel algorithms for natural language processing tasks. We now describe our results in this vein.

A key model class that we introduced is the recursive neural network [4]. Initially, we applied this to learning continuous phrase representations and syntactic parsing. Natural language parsing has typically been done with small sets of discrete categories such as NP and VP, but this representation does not capture the full syntactic nor semantic richness of linguistic phrases, and attempts to improve on this by lexicalizing phrases only partly address the problem at the cost of huge feature spaces and sparseness. To address this, we introduced a recursive neural network architecture for jointly parsing natural language and learning vector space representations for variable-sized inputs. At the core of our architecture are context-sensitive recursive neural networks (CRNN). These networks can induce distributed feature representations for unseen phrases and provide syntactic information to accurately predict phrase structure trees. Most

excitingly, the representation of each phrase also captures semantic information: For instance, the phrases "decline to comment" and "would not disclose the terms" are close by in the induced embedding space. This system achieves an unlabeled bracketing F-measure of 92.1% on the Wall Street Journal dataset for sentences up to length 15.

We next introduced a novel machine learning framework based on recursive autoencoders [6] for sentence-level prediction of sentiment label distributions. Our method learns vector space representations for multi-word phrases. In sentiment prediction tasks these representations outperform other state-of-the-art approaches on commonly used datasets, such as movie reviews, without using any pre-defined sentiment lexica or polarity shifting rules. We also evaluated the model's ability to predict sentiment distributions on a new dataset based on confessions from the experience project. The dataset consists of personal user stories annotated with multiple labels which, when aggregated, form a multinomial distribution that captures emotional reactions. Our algorithm can more accurately predict distributions over such labels compared to several competitive baselines.

We introduced our novel dynamic pooling algorithm [7] in the context of a paraphrase detection task. Paraphrase detection is the task of examining two sentences and determining whether they have the same meaning. In order to obtain high accuracy on this task, thorough syntactic and semantic analysis of the two statements is needed. We presented a method for paraphrase detection based on recursive autoencoders (RAE). Our unsupervised RAEs are based on a novel unfolding objective and learn feature vectors for phrases in syntactic trees. These features are used to measure the word- and phrase-wise similarity between two sentences. Since sentences may be of arbitrary length, the resulting matrix of similarity measures is of variable size. We developed a novel dynamic pooling layer which computes a fixed-sized representation from the variable-sized matrices. The pooled representation is then used as input to a classifier. Our method outperforms other state-of-the-art approaches on the challenging MSRP paraphrase corpus.

Based on the limitations of standard recursive neural networks, we introduced recursive matrix-vector spaces [11]. Single-word vector space models are very successful at learning lexical information. However, they cannot capture the compositional meaning of longer phrases, preventing them from a deeper understanding of language. We introduced a recursive neural network (RNN) model that learns compositional vector representations for phrases and sentences of arbitrary syntactic type and length. Our model assigns a vector and a matrix to every node in a parse tree: the vector captures the inherent meaning of the constituent, while the matrix captures how it changes the meaning of neighboring words or phrases. This matrix-vector RNN can learn the meaning of operators in propositional logic and natural language. The model obtains state of the art performance on three different experiments: predicting fine-grained sentiment distributions of adverb-adjective pairs; classifying sentiment labels of movie reviews and classifying semantic relationships such as cause-effect or topic-message between nouns using the syntactic path between them.

Knowledge bases provide applications with the benefit of easily accessible, systematic relational knowledge but often suffer in practice from their incompleteness and lack of knowledge of new entities and relations. Much work has focused on building or extending them by finding patterns in large unannotated text corpora. In contrast, we developed an algorithm to complete a knowledge base by predicting additional true relationships between entities, based on generalizations that can be discerned in the given knowledgebase. We introduced the neural tensor network (NTN) model [13] which predicts new relationship entries that can be added to the database. This model can be improved by initializing entity representations with word vectors learned in an unsupervised fashion from text, and when doing this, existing relations can even be queried for entities that were not present in the database. Our model generalizes and outperforms existing models for this problem, and can classify unseen relationships in WordNet with an accuracy of 75.8%.

Finally, we developed a novel approach to natural language parsing [22]. Natural language parsing has typically been done with small sets of discrete categories such as NP and VP, but this representation does not capture the full syntactic nor semantic richness of linguistic phrases, and attempts to improve on this by lexicalizing phrases or splitting categories only partly address the problem at the cost of huge feature spaces and sparseness. Instead, we introduce a Compositional Vector Grammar (CVG), which combines PCFGs with a syntactically untied recursive neural network that learns syntactico-semantic, compositional vector representations. The CVG improves the PCFG of the Stanford Parser by 3.8% to obtain an F1 score of 90.4%. It is fast to train and implemented approximately as an efficient reranker it is about 20% faster than the current Stanford factored parser. The CVG learns a soft notion of head words and improves performance on the types of ambiguities that require semantic information such as PP attachments.

Overall, our results in natural language processing have shown the promise of deep learning methods applied to these domains, and generated renewed interest in deep learning methods in the NLP community.

4.1.3 Speech recognition. While not a major focus of our efforts, we have also applied deep learning to speech recognition tasks. Work on deep neural networks as acoustic models for automatic speech recognition (ASR) have demonstrated substantial performance improvements. We introduced a model which uses a deep recurrent auto encoder neural network to denoise input features for robust ASR [23]. The model is trained on stereo (noisy and clean) audio features to predict clean features given noisy input. The model makes no assumptions about how noise affects the signal, nor the existence of distinct noise environments. Instead, the model can learn to model any type of distortion or additive noise given sufficient training data. The model is competitive with existing feature denoising approaches on the Aurora2 task, and outperforms a tandem approach where deep networks are used to predict phoneme posteriors directly.

4.1.4 Multimodal tasks. We evaluated a number of algorithms on tasks from multiple modalities, or on directly multimodal tasks.

While unsupervised feature learning is effective at learning representations that perform well on image, video and audio classification, many feature learning algorithms are hard to use and require extensive hyperparameter tuning. We developed sparse filtering [10], a simple new algorithm which is efficient and only has one hyperparameter, the number of features to learn. In contrast to most other feature learning methods, sparse filtering does not explicitly attempt to construct a model of the data distribution. Instead, it optimizes a simple cost function – the sparsity of l_2-normalized features – which can easily be implemented in a few lines of MATLAB code. Sparse filtering scales gracefully to handle high-dimensional inputs, and can also be used to learn meaningful features in additional layers with greedy layer-wise stacking. We evaluate sparse filtering on natural images, object classification (STL-10), and phone classification (TIMIT), and show that the method works well on a range of different modalities.

In another application, we focused on combining color and depth information. Advances in 3D sensing technologies make it possible to easily record color and depth images which together can improve object recognition. Most previous methods relied on very well-designed features for this new 3D modality. We introduced a model based on a combination of convolutional and recursive neural networks (CNN and RNN) for learning features and classifying RGB-D images [12]. The CNN layer learns low-level translationally invariant features which are then given as inputs to multiple, fixed-tree RNNs in order to compose higher order features. RNNs can be seen as combining convolution and pooling into one efficient, hierarchical operation. Our main result is that even RNNs with random weights compose powerful features. The model obtains state of the art performance on a standard RGB-D object dataset while being more accurate and faster during training and testing than comparable architectures such as two-layer CNNs.

Finally, we applied deep learning methods to audio/video multimodal tasks [24]. We presented a series of tasks for multimodal learning and showed how to train a deep network that learns features to address these tasks. In particular, we demonstrated cross modality feature learning, where better features for one modality (e.g., video) can be learned if multiple modalities (e.g., audio and video) are present at feature learning time. Furthermore, we showed how to learn a shared representation between modalities and evaluated it on a unique task, where the classifier is trained with audio-only data but tested with video-only data and vice-versa. We tested our methods on the CUAVE and AVLetters datasets with an audio-visual speech classification task, demonstrating superior visual speech classification on AVLetters and effective multimodal fusion.

Taken together, our results across a great variety of tasks demonstrates the potential of deep learning to address complex, multimodal data and successfully perform a variety of real-world perception tasks.

4.2 Results from achieving greater scale

A key hypothesis of our grant was that even simple models would have good performance if run at massive scale on big datasets. We detail results from this line of our

work in this section.

Our first effort in this area showed that even very simple unsupervised learning algorithms could attain high performance if run at scale [16]. A great deal of prior research focused on improving (typically, increasing the complexity of) algorithms for learning features from unlabeled data. And indeed, much progress was made on benchmark datasets like NORB and CIFAR by employing increasingly complex unsupervised learning algorithms and deep models. In our experiments, however, we showed that several very simple factors, such as the number of hidden nodes in the model, is as important to achieving high performance as the choice of learning algorithm or the depth of the model. Specifically, we applied several off-the-shelf feature learning algorithms (sparse auto-encoders, sparse RBMs and K-means clustering, Gaussian mixtures) to NORB and CIFAR datasets using only single-layer networks. We then performed a detailed analysis of the effect of changes in the model setup: the receptive field size, number of hidden nodes (features), the step-size ("stride") between extracted features, and the effect of whitening. Our results show that large numbers of hidden nodes and dense feature extraction are as critical to achieving high performance as the choice of algorithm itself—so critical, in fact, that when these parameters are pushed to their limits, we achieved state-of-the-art performance on both CIFAR and NORB using only a single layer of features. More surprisingly, our best performance is based on K-means clustering, which is extremely fast, has no hyper-parameters to tune beyond the model structure itself, and is very easy implement. Despite the simplicity of our system, we achieve performance beyond all previously published results on the CIFAR-10 and NORB datasets (79.6% and 97.0% accuracy respectively). The promising results from this initial study led us (and others) to the conclusion that scale matters, often even more than the learning algorithm or model.

To pursue greater scale, we developed new algorithms specifically aimed at overcoming barriers to scalability. For large deep network architectures the number of parameters can grow quadratically in the width of the network, thus necessitating hand-coded "local receptive fields" that limit the number of connections from lower level features to higher ones (e.g., based on spatial locality in images). To remove the need for such hand-coded modality-specific knowledge, we developed a fast method to choose these connections that may be incorporated into a wide variety of unsupervised training methods [8]. Specifically, we choose local receptive fields that group together those low-level features that are most similar to each other according to a pairwise similarity metric. This approach allows us to harness the advantages of local receptive fields (such as improved scalability, and reduced data requirements) when we do not know how to specify such receptive fields by hand or where our unsupervised training algorithm has no obvious generalization to a topographic setting. This method allowed us to use even simple unsupervised training algorithms to train successful multi-layered networks that achieved state-of-the-art results on CIFAR and STL datasets: 82.0% and 60.1% accuracy, respectively.

Another major barrier to scaling is the "whitening" preprocessing step common to a number of deep learning algorithms. One commonly used unsupervised learning

algorithm, Independent Component Analysis, requires an orthonormality constraint to be enforced, which makes it difficult to learn overcomplete features. In addition, ICA is sensitive to whitening. These properties make it challenging to scale ICA to high dimensional data. We developed a robust soft reconstruction cost for ICA that allows us to learn highly overcomplete sparse features even on unwhitened data [9]. Our formulation reveals formal connections between ICA and sparse autoencoders, which have previously been observed only empirically. Our algorithm can be used in conjunction with off-the-shelf fast unconstrained optimizers. We show that the soft reconstruction cost can also be used to prevent replicated features in tiled convolutional neural networks. Using our method to learn highly overcomplete sparse features and tiled convolutional neural networks, we obtained competitive performance on a wide variety of object recognition tasks, and state-of-the-art test accuracies on the STL-10 and Hollywood2 datasets.

We also investigated the best optimization approach in a large scale, distributed context. The predominant methodology in training deep learning advocates the use of stochastic gradient descent methods (SGDs). Despite its ease of implementation, SGDs are difficult to tune and parallelize. These problems make it challenging to develop, debug and scale up deep learning algorithms with SGDs. We showed that more sophisticated off-the-shelf optimization methods such as Limited memory BFGS (L-BFGS) and Conjugate gradient (CG) with line search can significantly simplify and speed up the process of pretraining deep networks [17]. In our experiments, the difference between LBFGS/CG and SGDs are more pronounced if we consider algorithmic extensions (e.g., sparsity regularization) and hardware extensions (e.g., GPUs or computer clusters). Our experiments with distributed optimization supported the use of L-BFGS with locally connected networks and convolutional neural networks. Using L-BFGS, our convolutional network model achieved 0.69% on the standard MNIST dataset, a state-of-the-art result among algorithms that do not use distortions or pretraining.

With these advances in hand, we attempted the first major attempt at massive-scale deep learning [14]. We considered the problem of building high-level, class-specific feature detectors from only unlabeled data. For example, is it possible to learn a face detector using only unlabeled images? To answer this, we trained a massive 9 layered locally connected sparse autoencoder with pooling and local contrast normalization on a large dataset of images (the model has 1 billion connections, the dataset has 10 million 200x200 pixel images downloaded from the Internet). We trained this network using model parallelism and asynchronous SGD on a cluster with 1,000 machines (16,000 cores) for three days. Our experimental results revealed that it is possible to train a face detector without having to label images as containing a face or not. Control experiments showed that this feature detector is robust not only to translation but also to scaling and out-of-plane rotation. We also found that the same network is sensitive to other high-level concepts such as cat faces and human bodies. Starting with these learned features, we trained our network to obtain 15.8% accuracy in recognizing 20,000 object categories from ImageNet, a leap of 70% relative improvement over the previous state-of-the-art. This result decisively demonstrated the benefits of scaling.

4.3 Theoretical results and control experiments

While most of our efforts were aimed at achieving better performance, we also conducted work aimed at a greater understanding of deep learning algorithms and their behavior. This section describes our efforts to understand—and ultimately improve—deep learning systems.

Two anomalous results in the literature demonstrated that certain feature learning architectures can perform very well on object recognition tasks, even without any training of the feature weights. In this theoretical work [18] we posed the question, why do random weights sometimes do so well? We proved that the answer lies in the architecture: certain convolutional pooling architectures are provably frequency selective and translation invariant, even with random weights. Based on this we demonstrated the viability of extremely fast architecture search by using random weights to evaluate candidate architectures, thereby sidestepping the time-consuming learning process. This work also showed that a surprising fraction of the performance of certain state-of-the-art methods can be attributed to the architecture alone, arguing against the learning algorithm per se as the site of performance differences between algorithms.

A vital element of deep learning and unsupervised feature learning is its domain general promise: a good algorithm might be expected to work without alteration on data from a variety of modalities like vision, audition, and natural language processing. We sought to directly evaluate the ability of a variety of unsupervised learning algorithms to learn good representations of different input data, by comparing the learned representations to those observed in the perceptual systems of biological organisms. We found that a number of unsupervised feature learning algorithms can account for features of normal neural receptive field properties across primary visual, auditory and somatosensory cortices [19]. Furthermore, we showed that the same algorithms explain the observed alterations in receptive field properties following experimental manipulation of an animal's early environment. These results make a contribution to theoretical neuroscience: Based on these modeling results we propose these models as phenomenological models of receptive field plasticity during an organism's lifetime. And due to the success of the same models in multiple sensory areas, we suggest that these algorithms may provide a constructive realization of the theory, first proposed by Mountcastle [1], that a qualitatively similar learning algorithm acts throughout primary sensory cortices. For the purposes of this grant, though, they revealed an important finding: all current deep learning algorithms fit biological data indistinguishably well. This is further evidence that the specific details of the learning algorithm matter less than the scale at which it is applied, and the architecture into which the learned features are placed.

Finally, in highly influential work, we directly investigated the relative importance of learning versus encoding [15]. While vector quantization (VQ) has been applied widely to generate features for visual recognition problems, much work has focused on more powerful methods. In particular, sparse coding has emerged as a strong alternative to traditional VQ approaches and has been shown to achieve consistently higher performance on benchmark datasets. Both approaches can be split into a training phase, where the system learns a dictionary of basis functions, and an encoding phase, where the

dictionary is used to extract features from new inputs. In this work, we investigate the reasons for the success of sparse coding over VQ by decoupling these phases, allowing us to separate out the contributions of training and encoding in a controlled way. Through extensive experiments on CIFAR, NORB and Caltech 101 datasets, we compare several training and encoding schemes, including sparse coding and a form of VQ with a soft threshold activation function. Our results show not only that we can use fast VQ algorithms for training, but that we can just as well use randomly chosen exemplars from the training set. Rather than spend resources on training, we find it is more important to choose a good encoder—which can often be a simple feed forward non-linearity. By choosing the best combination of learning algorithm and encoding algorithm, we obtained state-of-the-art performance on both the CIFAR and NORB object recognition datasets.

5.0 CONCLUSIONS

At the conclusion of this project, the goals outlined in the introduction stand fulfilled. A key finding of this report is that deep learning methods are capable of state-of-the-art performance across a huge variety of tasks in multiple modalities. This result has overturned key intuitions held across the machine learning, computer vision, speech recognition, and natural language processing communities: not only are hand-designed features no longer necessary, they are in fact now inferior to using deep learning methods. Researchers are thus now able to focus resources on modality-independent improvements, as better algorithms, greater scale, and faster optimization methods can be expected to improve state-of-the-art performance across a range of domains.

Based on our results from scaling algorithms up, we also conclude that pushing toward greater scale—both in model size and in dataset size—offers an important direction for increased performance. This trend towards "big data" and massive models, with very high capacity, trained in a purely supervised manner, is likely to continue to drive new applications in industry.

Our theoretical results all support the conclusion that, at present, the available unsupervised learning algorithms learn representations of equal quality. This overturned the widely-held intuition that more sophisticated (particularly probabilistic) learning methods were responsible for increasing performance in applications. We also conclude that current evaluation standards in the field, which focus on end-to-end system performance and permit simultaneous changes to many aspects (scale, learning algorithm, optimization method) of an algorithm, do not allow strong conclusions to be drawn as to the relative merits of single aspects of these algorithms. All that can be said is that an overall system combination is superior. Hence, we see a continued role for careful controlled experiments to tease apart the contributions of each design decision.

While many of the ideas and goals from the beginning of this project have proven durable throughout, certain parts of the project have evolved as our understanding has grown. Notably, at the outset of the project, we placed emphasis on using unlabeled data and unsupervised learning methods to reduce the amount of labeled data required for good performance. At the end of the project, this basic motivation remains intact—unsupervised deep learning methods are still best able to leverage very few labeled training examples—however, there has been a massive increase in available labeled data in many real world problems. This trend towards "big data" in industry has meant that interest has shifted from unsupervised learning to supervised learning in massive models. Unsupervised pretraining remains a useful regularizer to prevent overfitting in the regime where there are few labeled examples, but the challenge posed by "big data" is very different: the available datasets are so large that in fact underfitting is the main challenge. This trend has led to massive models, with very high capacity, trained in a purely supervised manner. Our recent work reflects this shift in focus.

The work arising from this project is continuing in a variety of paths. This project has supported the creation of a tutorial (the Stanford UFLDL Wiki), which has become a

widely used learning resource for those wishing to engage with deep learning. Students supported by this grant have gone on to positions at leading academic institutions (CMU; Princeton) and industry (Google; Baidu; Coursera). At the tail end of the grant, and continuing now, our lab has started a project aimed at vision-based autonomous driving. Leveraging massive labeled datasets collected on instrumented vehicles, this work aims to use deep learning at great scale to build low cost, low power, highly accurate perception systems for autonomous driving. It has been a great pleasure, and we are very grateful, to have contributed to the development of deep learning systems, now of use across a great variety of applications, which have taken another small step toward closing the gap between man and machine.

6.0 REFERENCES

1. **Tiled Convolutional Neural Networks.** Quoc V. Le, Jiquan Ngiam, Zhenghao Chen, Daniel Chia, Pangwei Koh and Andrew Y. Ng in NIPS 2010.

2. **Learning Hierarchical Spatio-Temporal Features for Action Recognition with Independent Subspace Analysis.** Quoc V. Le, Will Zou, Serena Yeung and Andrew Y. Ng in CVPR 2011.

3. **Learning Deep Energy Models.** Jiquan Ngiam, Zhenghao Chen, Pangwei Koh and Andrew Y. Ng in ICML 2011.

4. **Learning Continuous Phrase Representations and Syntactic Parsing with Recursive Neural Networks.** Richard Socher, Christopher Manning and Andrew Ng in NIPS 2010.

5. **Parsing Natural Scenes and Natural Language with Recursive Neural Networks.** Richard Socher, Cliff Lin, Andrew Y. Ng and Christopher Manning in ICML 2011.

6. **Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions.** Richard Socher, Jeffrey Pennington, Eric Huang, Andrew Y. Ng, and Christopher D. Manning in EMNLP 2011.

7. **Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection.** Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning in NIPS 2011.

8. **Selecting Receptive Fields in Deep Networks.** Adam Coates and Andrew Y. Ng in NIPS 2011.

9. **ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning.** Quoc V. Le, Alex Karpenko, Jiquan Ngiam and Andrew Y. Ng in NIPS 2011**.**

10. **Sparse Filtering.** Jiquan Ngiam, Pangwei Koh, Zhenghao Chen, Sonia Bhaskar and Andrew Y. Ng in NIPS 2011.

11. **Semantic Compositionality through Recursive Matrix-Vector Spaces.** Richard Socher, Brody Huval, Christopher D. Manning and Andrew Y. Ng in EMNLP 2012.

12. **Convolutional-Recursive Deep Learning for 3D Object Classification.** Richard Socher, Brody Huval, Bharath Bhat, Christopher D. Manning, Andrew Y. Ng in NIPS 2012.

13. **Learning New Facts From Knowledge Bases With Neural Tensor Networks and Semantic Word Vectors.** Danqi Chen, Richard Socher, Christopher D. Manning, Andrew Y. Ng in ICLR 2013.

14. **Building High-Level Features using Large Scale Unsupervised Learning.** Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeffrey Dean and Andrew Y. Ng in ICML 2012.

15. **The Importance of Encoding Versus Training with Sparse Coding and Vector Quantization.** Adam Coates and Andrew Y. Ng in ICML 2011.

16. **An Analysis of Single-Layer Networks in Unsupervised Feature Learning.** Adam Coates, Honglak Lee and Andrew Ng in AISTATS 14, 2011.

17. **On Optimization Methods for Deep Learning.** Quoc V. Le, Jiquan Ngiam, Adam Coates, Abhik Lahiri, Bobby Prochnow and Andrew Y. Ng in ICML 2011.

18. **On Random Weights and Unsupervised Feature Learning.** Andrew Saxe, Pangwei Koh, Zhenghao Chen, Maneesh Bhand, Bipin Suresh and Andrew Y. Ng in ICML 2011.

19. **Unsupervised Learning Models of Primary Cortical Receptive Fields and Receptive Field Plasticity.** Andrew Saxe, Maneesh Bhand, Ritvik Mudur, Bipin Suresh and Andrew Y. Ng in NIPS 2011.

20. **End-to-End Text Recognition with Convolutional Neural Networks.** Tao Wang, David J. Wu, Adam Coates and Andrew Y. Ng in ICPR 2012.

21. **Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning.** Adam Coates, Blake Carpenter, Carl Case, Sanjeev Satheesh, Bipin Suresh, Tao Wang, David Wu and Andrew Y. Ng in ICDAR 2011.

22. **Parsing with Compositional Vector Grammars.** John Bauer, Richard Socher, Christopher D. Manning, Andrew Y. Ng in ACL 2013.

23. **Recurrent Neural Networks for Noise Reduction in Robust ASR.** A.L. Maas, Q.V. Le, T.M. O'Neil, O. Vinyals, P. Nguyen, and Andrew Y. Ng in Interspeech 2012.

24. **Multimodal Deep Learning.** Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee and Andrew Y. Ng in ICML 2011.

List of Acronyms

ASR	Automatic Speech Recognition
CNN	Convolutional Neural Networks
COTS	Commercial Off The Shelf
CRNN	Context-sensitive Recursive Neural Networks
CVG	Compositional Vector Grammar
GPS	Graphics Processing Unit
HOG	Histogram of Oriented Gradients
ICA	Independent Component Analysis
ISA	IndependentSubspace Analysis
NLP	Natural Language Parsing
NORB	NYU Object Recognition Benchmark
NP	Noun Phrases
NTN	Neural Tensor Network
PCFG	Probabilistic Context Free Grammar
PP	Prepositional Phrases
RAE	Recursive AutoEncoders
RBM	Recursive Boltzman Machine
RICA	Reconstruction Independent Components Analysis
RNN	Recursive Neural Networks
SIFT	Scale Invariant Feature Transform
VP	Verb Phrases
CIFAR-10	Canadian Institute for Advanced Research funded dataset which has 10 classes of objects with 6000 images each
Hollywood2	Action dataset based on movie excerpts